Curiosity-driven phonetic learning

Clément Moulin-Frier, Pierre-Yves Oudeyer

Curiosity-driven phonetic learning

Clément Moulin-Frier
INRIA Bordeaux Sud-Ouest
Ensta-ParisTech
France
Email: clement.moulin-frier@inria.fr

Pierre-Yves Oudeyer
INRIA Bordeaux Sud-Ouest
Ensta-ParisTech
France
Email: pierre-yves.oudeyer@inria.fr

***Abstract*—This article studies how developmental phonetic learning can be guided by pure curiosity-driven exploration, also called intrinsically motivated exploration. Phonetic learning refers here to learning how to control a vocal tract to reach acoustic goals. We compare three different exploration strategies for learning the auditory-motor inverse model: random motor exploration, random goal selection with reaching, and curiosity-driven active goal selection with reaching. Using a realistic vocal tract model, we show how intrinsically motivated learning driven by competence progress can generate automatically developmental structure in both articulatory and auditory modalities, displaying patterns in line with some experimental data from infants.**

I. Introduction

In their first months, human infants spontaneously explore how to produce vocalizations, learning the mapping between motor commands controlling the vocal tract and their acoustic consequences [1]. We study here, in a simulated robotic setup, how various strategies of spontaneous exploration, including intrinsically motivated exploration, can generate developmental structures in early vocal learning while allowing a robotic speaker to learn its auditory-motor inverse model. Speech production general principles are illustrated in Figure 1.

Let us mention two major works in the field of computational models of human vocal learning, which we also refer as *phonetic learning* (although these models extend to higher linguistic levels). The DIVA model [2], [3] proposes an architecture partly inspired by neurolinguistics. It involves two learning phases. The first one is analogous to infant babbling and corresponds to semi-random articulator movements producing auditory and somatosensory feedbacks. This is used to tune a neural network between representation maps. In the second phase, the model is exposed to external speech sounds analogous to an ambient language and learn how to produce them adequately. The Elija model [4] also distinguishes several learning phases. The phases related to phonetic learning are driven by a reward function (including sound salience and diversity, as well as articulatory effort). The sounds produced by the model then attract the attention of a caregiver, thus providing an external reinforcement signal.

By focusing on phonetic learning, our study is limited to the first learning phase in DIVA and Elija, in which the former involves a semi-random articulatory exploration and the latter a hand-coded reward function. Rather than considering a pre-determined exploration, we are interested in the internal mechanisms which can drive adaptive phonetic exploration and learning instead and in the early stage of spontaneous vocal exploration.

This process of vocal learning is here framed as an instance of a more general robotics motor learning problem, that of learning the inverse model mapping a distribution of perceptual effects to the corresponding distribution of motor programs that generate these effects [5]. Phonetic learning shares several fundamental properties with learning other kinds of inverse models like inverse body kinematics, visual hand reaching, or locomotion: the corresponding sensorimotor spaces are high-dimensional, highly redundant and non-linear, and too large to be explored and learnt entirely in a lifetime. In previous work about such inverse model learning, we have shown the importance of developmental mechanisms guiding exploration and learning in these spaces [5], [6]. Among these guiding mechanisms, intrinsic motivations, generating spontaneous exploration in humans [7], [8], have been transposed in curiosity-driven learning machines [9]–[11] and robots [5], [6] and shown to yield highly efficient learning of inverse models in high-dimensional redundant sensorimotor spaces [5]. Efficient versions of such mechanisms are based on the active choice of learning experiments that maximize learning *progress*, for e.g. improvement of predictions or of competences to reach goals [6], [9]. This automatically drives the system to explore and learn first easy skills, and then explore skills of progressively increasing complexity. Such intrinsically motivated exploration was also shown to generate automatically behavioural and cognitive developmental structures sharing interesting similarities with infant development [6], [12], [13]. This approach is grounded in psychological theories of intrinsic motivations [7], [14], explores several fundamental questions about curiosity-driven open-ended learning in robots [6], and allows to generate some novel hypotheses for the explanation of infant development, regarding behavioural [13], cognitive [12] and brain circuitry [15].

Additionally, it was shown in previous models that learning redundant inverse models could be achieved more efficiently if exploration was driven by goal babbling, triggering reaching, rather than driven by direct motor babbling [5], [16].

We thus explore here how phonetic learning can be achieved with intrinsically motivated goal exploration mechanisms, and what developmental structure it may generate. In an experi-

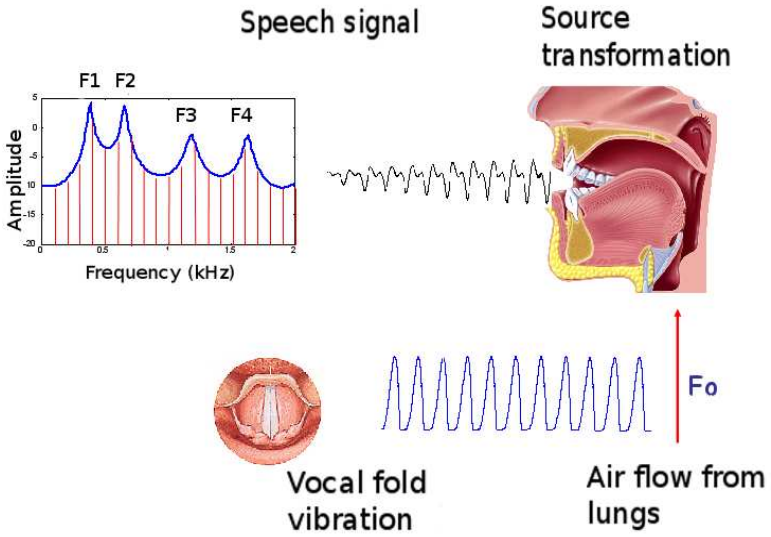


Fig. 1. Speech production general principles. The vocal fold vibration by the lung air flow provides a source signal: a complex sound wave with fundamental frequency F_0. According to the vocal tract shape, acting as a resonator, the harmonics of the source fundamental frequency are selectively amplified or faded. The local maxima of the resulting spectrum are called the formants, ordered from the lower to the higher frequency. They belong to the major features of speech perception.

mental setup using the VLAM model of vocal production [17], we compare such a strategy with random motor babbling as well as with random goal babbling.

II. Exploration strategies

We consider an agent provided with a motor space M corresponding to articulatory commands controlling the shape of its vocal tract, and a sensory space S corresponding to acoustic features. Both spaces are continuous. The agent, which has to be considered in its initial state as a pre-vocalizing baby agent, does not know any relationship between these variables. Let us call $f : M \rightarrow S$ the function defined by the physical properties of the environment (mainly aero-acoustic laws). The aim of phonetic learning is to approximate f^{-1}, that is the inverse mapping from acoustic goals to reach in S, to adequate motor commands in M. To do this, the agent can observe $(m, s) \in M \times S$ pairs from its own experience, and thus has to deal with three main issues:

- M and S can be highly dimensional, such that random sampling to collect (m, s) pairs would lead to too sparse data for an efficient learning;
- f can be strongly non-linear, such that function approximation from experience is not trivial;
- f can be redundant (many M to one S), such that f^{-1} approximation is a ill-posed problem.

When a learning process faces these three issues, as it is the case in phonetic learning, random exploration in M is not a realist strategy to collect (m, s) pairs. Due to high dimensionality, data are precious whereas, due to non-linearity and/or redundancy, data are not equally useful to learn the inverse mapping f^{-1}. As collecting a (m, s) pair involves the realization of m, through f, to observe s, the problem is then to find good learning strategies. Let us consider three different ones.

- **Random motor exploration:** at each time step, the agent randomly chooses an articulatory command $m \in M$, produces it, and observes $s = f(m)$.
- **Random goal selection with reaching:** at each time step, the agent randomly chooses a goal $s_g \in S$ and tries to reach it by producing $\{m_1, \ldots, m_n\} \in M^n$. It observes the corresponding sensory consequences $\{s_1, \ldots, s_n\} \in S^n$.
- **Active goal selection with reaching:** at each time step, the agent choose a goal s_g according to a measure of interest in S based on its previous experiences. It tries to reach s_g by producing $\{m_1, \ldots, m_n\} \in M^n$ and observes the corresponding sensory consequences $\{s_1, \ldots, s_n\} \in S^n$. It updates the interest measure with respect to these new experiences.

These three learning strategies are similar to the ACTUATOR-RANDOM, SAGG-RANDOM and SAGG-RIAC algorithms in [5], respectively. They differ in two ways. Firstly by what we call the *choice space*, which refer to the space in which the initial point is drawn in order to collect a (m, s) pair. When the choice space is M, as it is the case for the random exploration strategy, the agent can directly produced an m and thus collect a $(m, s = f(m))$ pair. When it is S, as it is the case in the random and active goal selection strategies, the agent has to find a way to reach this goal choice. It typically consists in an optimization procedure which requires several trials (either by actual realizations through f or alternatively using a model based on previously collected (m, s) pairs). Secondly, they differ by an active or random selection in the choice space. When this latter is M,

we only consider random selection (although active selection can be conceived, see [5]). When it is S, active selection refers to the ability of actively choosing a goal with respect to a measure of interest $I : S \rightarrow \mathbb{R}$. In previous papers (eg [18]), we showed that an adequate interest measure is the *competence progress*. It is computed from an history of previous competences in reaching goals in regions of the choice space.

III. Developmental robotics experiment

This section first describes the vocal tract model we use in our experiments, then exposes a specific implementation of the learning strategies proposed above.

A. Articulatory-acoustic model

The function f defining the articulatory-to-acoustic transformation corresponds to a vocal tract model able to compute the sound wave generated by a given articulatory configuration.

We use a realistic system modeling the vocal tract, the Variable Linear Articulatory Model [19] derived from the Maeda articulatory model [20]. This latter was conceived from a statistical analysis of 519 vocal tract sagittal contours from radiographic measurements and tomographic studies of sentences pronounced by a French speaker. These contours are segmented in 28 sections from the glottis to the lips, from which the corresponding vocal tract areas are calculated. This analysis provides seven main parameters explaining 88% of the data variance, and which may be interpreted in terms of phonetic commands corresponding to the jaw (J), the tongue body (TB), dorsum (TD) and tip (TT), the lip protrusion (LP) and separation height (LH), as well as the larynx height (Lx) (Figure 2). These parameters can in turns be linearly combined to reconstruct the sagittal contour and then the area function. The formants and the transfer function are finally calculated from this latter, and a sound can be generated from formant frequencies and bandwidths (Figure 3). In other words, VLAM models the speech production process depicted Figure 1. In this experiment, we limit ourselves to vowel generation where the vocal tract has to be sufficiently open (minimum of the area function greater than 0.15 cm^2).

VLAM inputs and outputs should then be transformed into adequate representations (see [21] for a discussion about realistic perceptual and motor representations of speech gestures in an articulatory model). For the articulatory space M, we use the seven VLAM parameters (Figure III-A). Note that in VLAM the articulatory space is centered such that the neutral position corresponds to null values of the parameters. For the auditory space S, we use a common two-dimensional representation for vowels [22]–[24]. The first dimension is the first formant F_1. The second one is the second effective formants F_2', which correspond to a weighted average of F_2, F_3 and F_3. They are are expressed in Barks [25], a psychoacoustic measurement reflecting human frequency perception. Figure 4 display the space S, usually called the vocalic triangle.

Algorithm 1 Reaching phase algorithm for goal selection strategies. $s_g \in S$ is the goal to reach. $\mathcal{N}(\mu, \sigma)$ is the multivariate normal distribution with mean μ and standard deviation σ on each dimension.

$s_{best} \leftarrow NN(s_g)$
$m_{best} \leftarrow M(s_{best})$

for N trials **do**
  $m \sim \mathcal{N}(m_{best}, \sigma_{expl})$
  $s \leftarrow f(m)$
  if $\|s_g - s\| < \|s_g - s_{best}\|$ **then**
    $m_{best} \leftarrow m$
    $s_{best} \leftarrow s$
  end if
end for

B. Learning strategy implementation

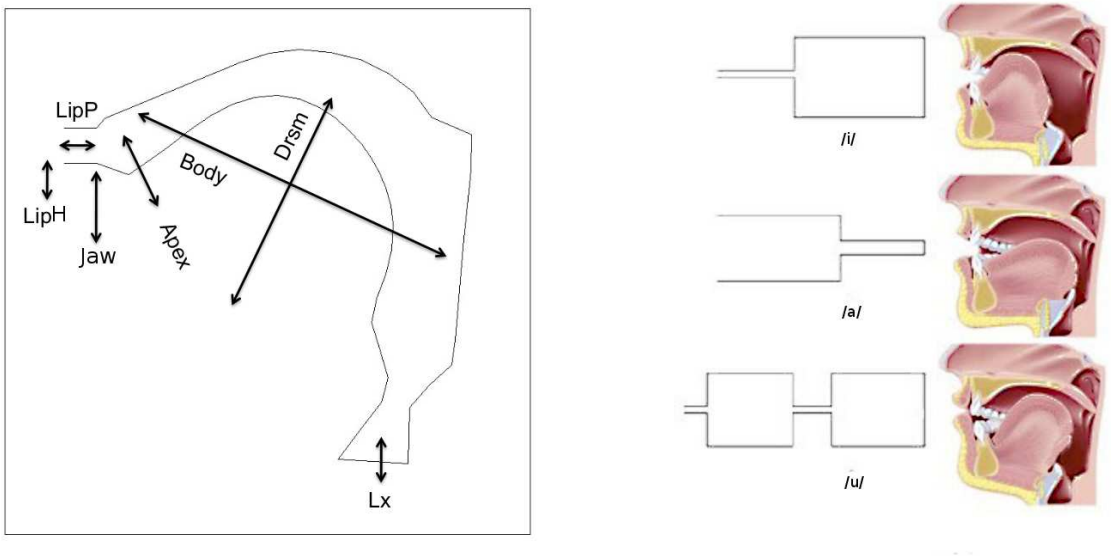
We assume that the agent is provided with an episodic memory of the previously experienced (m, s) pairs. Given a request $s_g \in S$, it is able to find the (m_i, s_i) pair in its memory which minimize the distance $\|s_i - s_g\|$. In other words, the memory provides a link between M and S, as well as a nearest neighbor search procedure in S. We note $M(s)$ the m associated with a s in a (m, s) pair and $NN(s)$ the nearest neighbor of s in the memory.

1) Random motor exploration: Random motor exploration is the simplest learning strategy we consider. It consists of successively drawing motor configurations in M according to a uniform distribution.

2) Random goal selection with reaching: Instead of using the motor space M as the choice space, as defined in the previous section, this learning strategy uses the sensory space S. Once a particular goal s_g has been drawn from an uniform distribution over S, it requires a subsequent reaching phase in which the agent has to perform an optimization procedure to provide an adequate motor command m in order to reach the goal s_g. We choose a simple procedure based on a mutation-selection loop as described in Algorithm 1.

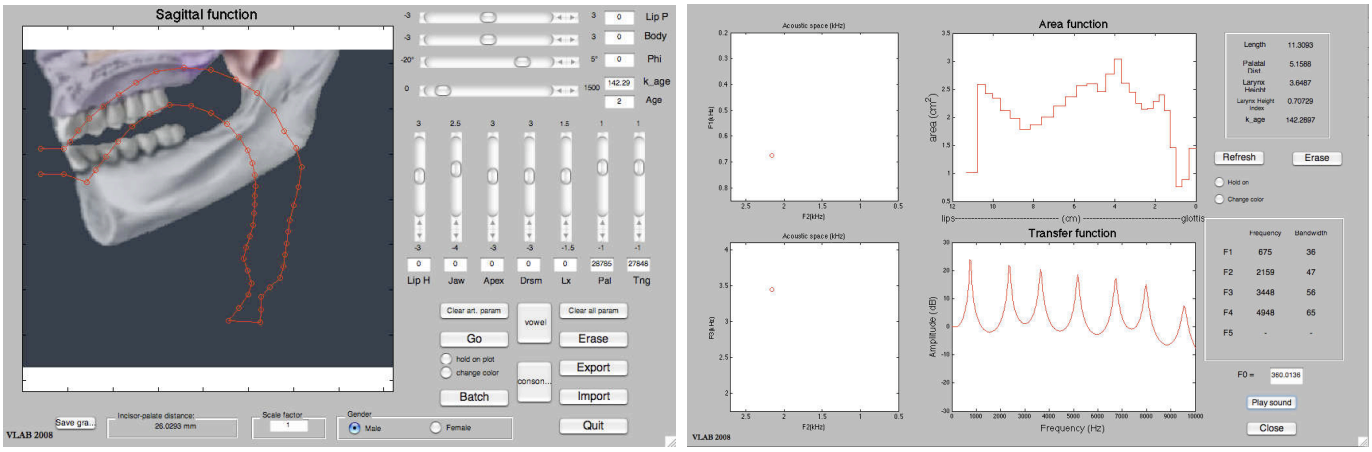
This implementation of reaching requires at least one pre-existing (m, s) pair in order to find a nearest neighbor for the first goal. We choose to bootstrap the system by producing articulatory commands close to a neutral position during the first 100 experiences. This local reaching procedure can therefore be conceived as a maturational-like mechanism, starting on a neutral position and then exploring variations around what was already tried.

3) Active goal selection with active reaching: This learning strategy also uses S as the choice space but involves an active goal selection based on a competence progress measure. Before running the reaching phase defined above, the agent looks at the nearest neighbor $s = NN(s_g)$ in its past sensory experiences. Once the reaching phase is performed, it computes the difference $d = \|s_g - s\| - \|s_g - s_{best}\|$. This is used to compute the competence progress $cp(s_g)$ over the S space, which will act as a measure of interest to



(a) (b)

Fig. 2. (a): The seven articulatory parameters in VLAM. The lips are at the left, the vocal folds at the bottom right. The jaw controls the global opening of the vocal tract. The apex controls the the tongue tip position. Body and dorsum globally control the front/back and low/high dimensions of the tongue, respectively. Lx controls the larynx height. (b): Some common configurations (right) together with a tube representation representing the global shape of the oral cavity (left). This latter can be viewed as an approximation of the area function. /i/ corresponds to a tight constriction at the front of the vocal tract; /a/ to a wider constriction at the back, and /u/ to a tight constriction at the middle of the vocal tract and at the lips.



(a) Articulatory part (b) Acoustic part

Fig. 3. VLAM processing flow. Articulatory part: a vocal tract shape is generated from the seven articulatory commands; Acoustic part: from the area function (top right), the spectrum of the vocal tract transfer function is computed (bottom right) leading to formant values positioned in the (F_1, F_2) and (F_2, F_3) spaces (left).

select subsequent goals. We define the competence progress as $e^{-d} \in [0, 1]$, where d is always non-negative due to the reaching implementation of Algorithm 1. Thus, the gain in reaching distances is emphasized for goals close to be reached, allowing to focus on the reachable parts of S. We nevertheless define a threshold ϵ for which goals are considered to be reached, such that:

$$\forall d \leq \epsilon : \ cp(d) = cp(0) = 1,$$
$$\forall d > \epsilon : \ cp(d) = e^{-d}.$$

Technically, the goal space is discretized in a fine-grained grid over S (generally 50 bins by dimension), in which a time-weighted measure of the competence progress is maintained. Each cell i of the grid starts with a null competence progress value $CP_i(t = 0) = 0$. Then, each time t a goal s_g is selected and leads to a competence progress value $cp(s_g)$, the

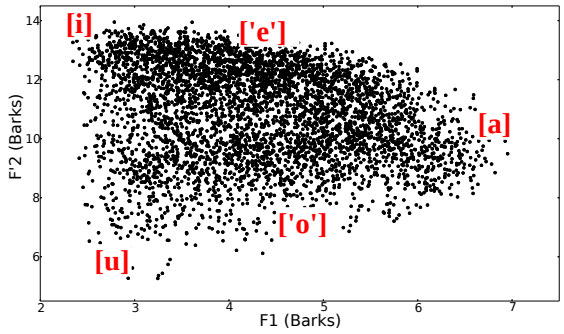


Fig. 4. Auditory space in (F_1, F_2') from a uniform sampling of the VLAM 7-dimensional articulatory space, with common phone locations. Globally, F_1 is rather correlated with the constriction size and F_2' with the constriction position (see Figure 2b).

corresponding cell is updated with the following formula:

$$CP_i(t) = \begin{cases} \alpha.cp(s_g) + (1-\alpha)CP_i(t-1) & \text{if } s_g \text{ in cell } i \\ CP_i(t-1) & \text{otherwise.} \end{cases} \quad (1)$$

This allows a fading memory of the competence progress measure in each cell. Generalization across cells is done by a gaussian filtering with standard deviation expressed in number of bins (generally 4).

IV. Results

A. Random exploration

Figure 5 shows the distribution of produced sounds over time for the random exploration strategy. We observe that it does not evolve over time, as they are always produced from a random drawing in M space.

B. Random goal selection with reaching

Figure 6 shows the distribution of produced sounds over time for the random goal selection with reaching strategy. During its first 100 vocalizations, the agent is in a bootstrap mode such that motor commands are around the neutral position, leading to sensory consequences concentrated on a small part of S. Then, we observe a progressive exploring of the sensory space S. This is due to the conjoint action of random goal selection (which push to cover the whole space) and reaching by local exploration (which provide the progressive aspect).

C. Active goal selection with reaching

Figure 7 shows the distribution of produced sound over time for the active goal selection with reaching strategy. At the beginning, we observe a similar behavior compared to the random goal selection strategy. However, it covers more uniformly the sensory space S at the end. This is due to the active goal selection which pushes him to focus on regions which maximize the competence progress, allowing to adapt its focus on less visited parts of S.

This is somewhat in line with some developmental observations [1] showing a tendency for childrens to begin producing front vowels (with high F_2 values) and progressively shift back their tongue during maturation (leading to back vowels with low F_2 values).

Figure 8 shows the distribution of the produced articulatory commands in M over time. Firstly, we indeed observe a progressive shift of the tongue body TB (front-back dimension of the tongue). Secondly, we observe that some structure emerges, in the sense that some articulators are preferred. The random goal selection strategy exposes a very similar behavior (not shown here). To interpret this articulatory pattern, let us consider the extreme case where an articulator does not have any influence on the produced sound. According to the reaching algorithm in Algorithm 1, it should then describe a Brownian motion, thus resulting in a bell-shaped distribution centered around the neutral position. Goal selection with reaching phase strategies therefore mainly use articulators playing an important role in goal reaching.

However, the jaw J and the tongue dorsum TD does not seem to be much involved whereas they normally play an important role, both controlling the size of the vocal tract constriction. This latter plays a major role to control the F_1 value (tight constriction leads to low F_1 values as in /i/, and large one to high F_1 values as in /a/). Figure 9 shows the conjoint density of these two articulators. We observe a rather strong correlation indicating a conjoint action of both articulators, thus explaining their rather low respective mobility in Figure 8.

Interestingly, the used articulators correspond to the minimal set allowing adequate production of vowels [26], although the jaw is poorly involved whereas it is the main articulator of vocal babbling. This suggests that further extensions using dynamical articulator trajectories are needed to relate the model to more experimental data.

To conclude this result section, note that we do not provide score comparison of the learning strategies on a control task. Actually, at the present state of our modeling process, there is a tradeoff between good comparison results (showing globally that choosing in S is better than in M, and that an active selection is better than a random one, see [5] for thorough results on score comparison), and interesting resulting developmental patterns. These parameters are mainly the standard deviation around the neutral position used to bootstrap goal selection strategies, and the standard deviation σ_{expl} in the reaching Algorithm 1. We deliberately choose very small values for both parameters to focus on developmental sequence observations.

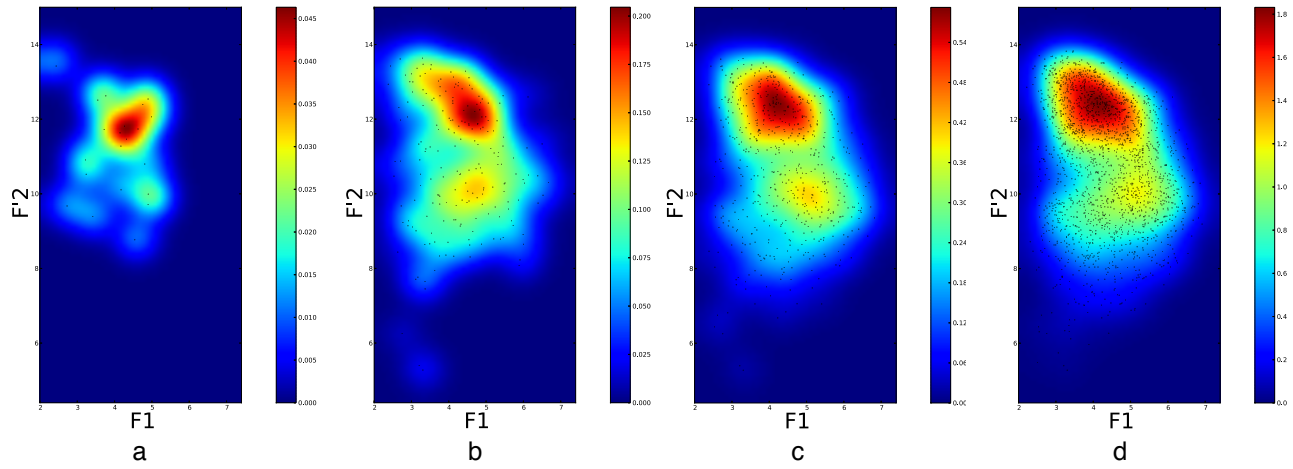


Fig. 5. Densities of produced sounds over the space S in the random exploration strategy. a) after 100 vocalizations; b) after 1000 vocalizations; c) after 3000 vocalizations; d) after 10000 vocalizations. Formants are expressed in Barks. Color bars are expressed in number of produced sounds per bins, with 50 bins per dimension.

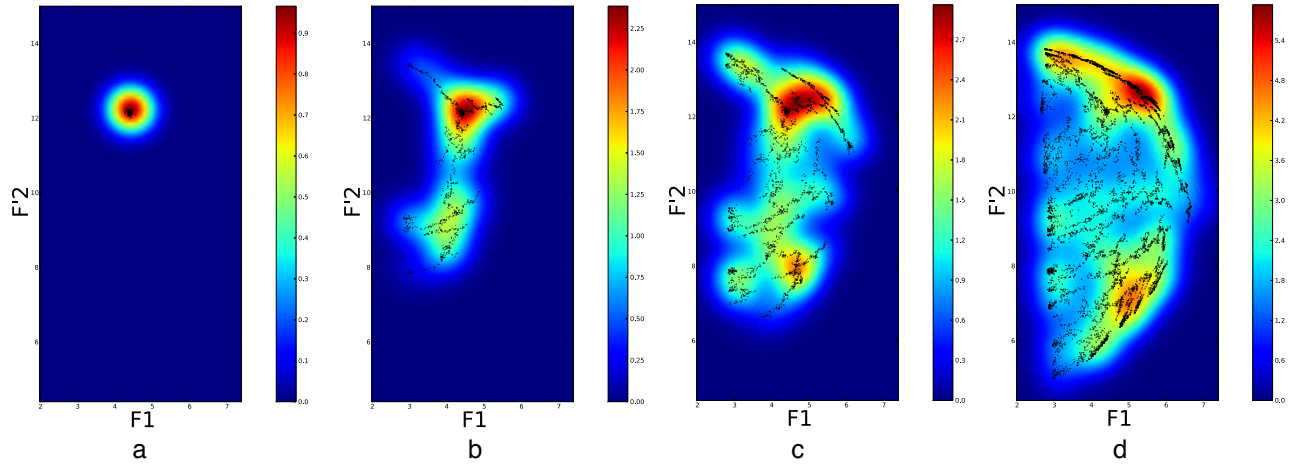


Fig. 6. Densities of produced sounds over the space S in the random goal selection with reaching strategy. Same convention than in Figure 5, the number of vocalizations also including those performed during the reaching phase.

V. Conclusion

This preliminary study showed how goal-directed learning driven by the competence progress can let emerge a developmental structure in the articulatory and acoustic spaces. We showed interesting developmental patterns in both spaces. Firstly, active goal selection displays a progressive exploration of the auditory space, from sensory consequences of a neutral articulatory configuration to a whole covering of S, which is relatively in line with some experimental data. Secondly, the articulators seem to be recruited according to their respective efficiency to reach goals in the acoustic space.

These results thus encourage further extensions of the model. More specifically, this research project aims at proposing an integrated computational model of language acquisition based on the interaction of three subsystems:

- an intrinsic motivation system allowing the agent to focus on goals which maximize the competence progress. Further extensions would involve higher levels of goals, for example related to the use of vocalization to denote external referents, exploring the path toward semantics.
- a social guidance system allowing the agent to be influenced by an external skilled agent and providing goal suggestions and/or action demonstrations, either by a human or by another robotic agent.
- a maturational system allowing the agent to progressively release its sensory-motor constraints according to its competence progress, through motor primitives encoding dynamical articulator properties (e.g. [27]).

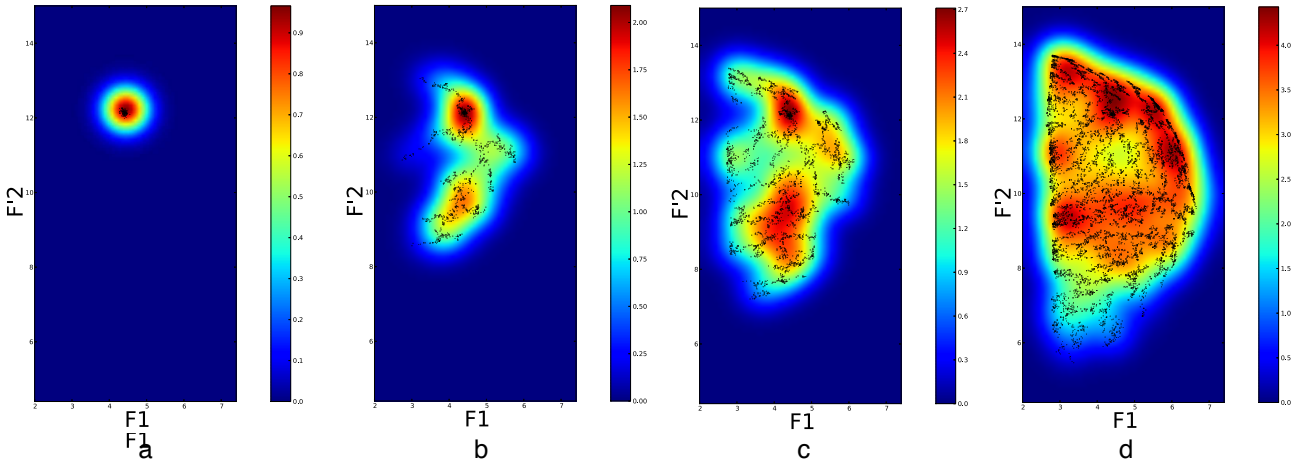


Fig. 7. Densities of produced sounds over the space S in the active goal selection with reaching strategy. Same convention than in Figure 5, the number of vocalizations also including those performed during the reaching phase.

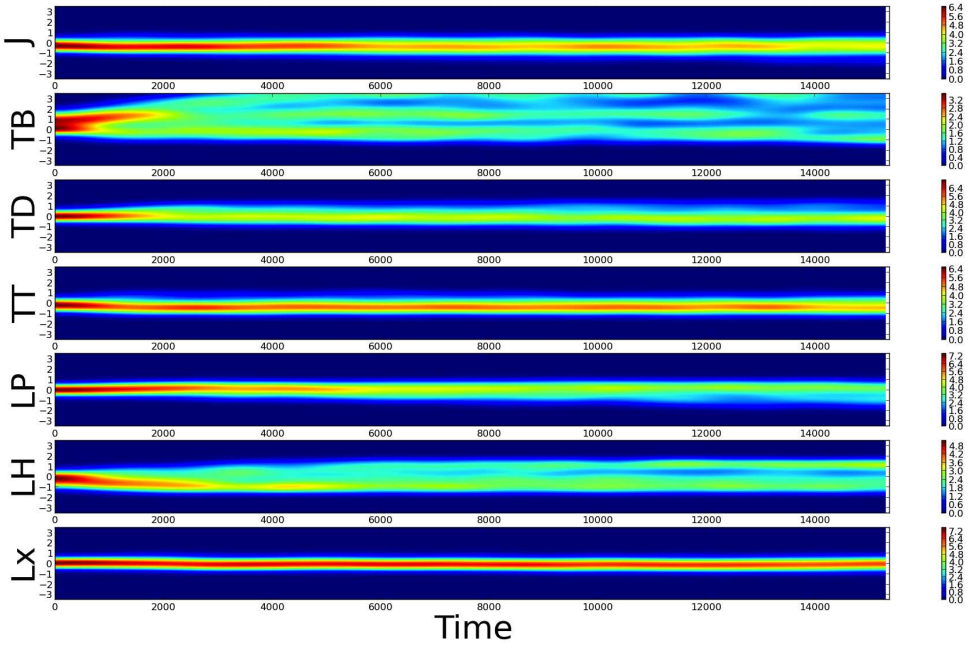


Fig. 8. Densities over time of produced articulatory commands over the space M in the active goal selection with reaching strategy. 0 values on the y-axis corresponds to the neutral position in VLAM. High TB values correspond to a tongue back in the mouth; high LH values to open lips. Other articulators do not need orientation information as they relatively stay around their neutral positions.

We also want to apply this approach to the control of a more complex articulatory synthesizer. We are interested in using the free software Praat [28], a powerful tool allowing to synthesize a speech signal from a trajectory in a 29-dimensional space of respiratory and oro-facial muscles. Numerous acoustic features can in turn be extracted from the synthesized sound, among which the Mel-frequency cepstral coefficients (MFCC, [29]). Our hope is that a developmental robotics approach applied to a realistic articulatory model can appropriately manage the learning process of this complex mapping in high-dimensional spaces , and that observed developmental sequences can lead to interesting experimental data comparisons and predictions. In particular, using such a dynamic model controlled by muscle activity could hopefully allow to relate our results to more common speech acquisition data, in particular regarding sub-glottal exploration and babbling.

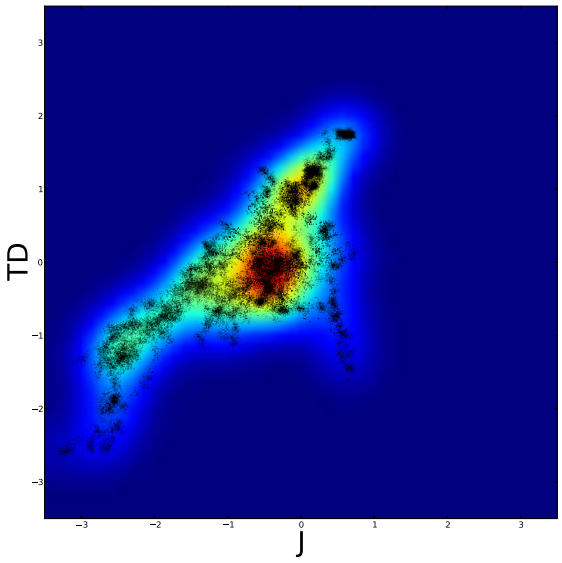


Fig. 9. Conjoint density over the jaw (J) and tongue dorsum (TD) articulators. Active goal selection with reaching strategy.

ACKNOWLEDGMENT

This work was partially financed by ERC Starting Grant EXPLORERS 240 007.

The VLAM model was developed by Shinji Maeda, then modified and integrated into a speech synthesis environment called VLAB by Louis-Jean Boë and a team from GIPSA-lab at Grenoble University (Jean-Luc Schwartz, Pierre Badin, Laurent Girin, Frédéric Berthommier and numerous studients from the IUT d'informatique de Grenoble), in particular in the framework of the SkullSpeech project (ANR, CNRS).

We would also like to thank Fabien Benureau for crucial help around the time of the deadline.

REFERENCES

[1] D. K. Oller, *The emergence of the sounds of speech in infancy*. Academic Press, 1980, vol. 1, ch. 6, pp. 93–112.

[2] F. H. Guenther, M. Hampson, and D. Johnson, "A theoretical investigation of reference frames for the planning of speech movements," *Psychological Review*, vol. 105, no. 4, pp. 611–633, Oct. 1998, PMID: 9830375.

[3] F. H. Guenther, "Cortical interactions underlying the production of speech sounds," *Journal of Communication Disorders*, vol. 39, no. 5, pp. 350–365, Sep. 2006.

[4] I. Howard and P. Messum, "Modeling the development of pronunciation in infant speech acquisition," *Motor Control*, vol. 15(1), pp. 85–117, 2011.

[5] A. Baranes and P.-Y. Oudeyer, "Active learning of inverse models with intrinsically motivated goal exploration in robots," *Robotics and Autonomous Systems*, in press.

[6] P.-Y. Oudeyer, F. Kaplan, and V. Hafner, "Intrinsic motivation systems for autonomous mental development," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 2, pp. 265–286, 2007.

[7] D. E. Berlyne, "A theory of human curiosity," *British Journal of Psychology*, vol. 45, pp. 180–191, 1954.

[8] E. Deci and R. M. Ryan, *Intrinsic Motivation and self-determination in human behavior*. New York: Plenum Press, 1985.

[9] J. Schmidhuber, "A possibility for implementing curiosity and boredom in model-building neural controllers," in *Proc. SAB'91*, J. A. Meyer and S. W. Wilson, Eds., 1991, pp. 222–227.

[10] A. Barto, S. Singh, and N. Chenatez, "Intrinsically motivated learning of hierarchical collections of skills," in *Proc. 3rd Int. Conf. Dvp. Learn.*, San Diego, CA, 2004, pp. 112–119.

[11] J. Schmidhuber, "Formal theory of creativity, fun, and intrinsic motivation (1990-2010)," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 3, pp. 230–247, 2010.

[12] F. Kaplan and P.-Y. Oudeyer, "The progress-drive hypothesis: an interpretation of early imitation," in *Models and mechanisms of imitation and social learning: Behavioural, social and communication dimensions*, K. Dautenhahn and C. Nehaniv, Eds. Cambridge University Press, 2005.

[13] P.-Y. Oudeyer and F. Kaplan, "Discovering communication," *Connection Science*, vol. 18, no. 2, pp. 189–206, 2006.

[14] M. Csikszentmihalyi, *Creativity: Flow and the Psychology of Discovery and Invention*. HarperCollins, 1997. [Online]. Available: http://books.google.fr/books?id=aci_Ea4c6woC

[15] F. Kaplan and P.-Y. Oudeyer, "In search of the neural circuits of intrinsic motivation," *Frontiers in Neuroscience*, vol. 1, no. 17, 2007.

[16] M. Rolf, J. Steil, and M. Gienger, "Online goal babbling for rapid bootstrapping of inverse models in high dimensions," in *Proceeding of the IEEE ICDL-EpiRob Joint Conference (2011)*, 2011.

[17] L. Boë, N. Valle, P. Badin, J. Schwartz, and C. Abry, "Tendencies in phonological structures: the influence of substance on form," *Bulletin de la Communication Parle*, vol. 5, pp. 35–55, 2000.

[18] P.-Y. Oudeyer and F. Kaplan, "Discovering communication," *Connection Science*, vol. 18, no. 2, pp. 189–206, 06 2006.

[19] L. Boë, "Vowel spaces of newly-born infants and adults consequences for ontogenesis and phylogenesis," in *14th Internation Congress of Phonetic Sciences*, 1999, pp. 2501–2504.

[20] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model," *Speech production and speech modelling*, pp. 131–149, 1989.

[21] J. Serkhane, J.-L. Schwartz, and P. Bessiere, "Building a talking baby robot: A contribution to the study of speech acquisition and evolution," *Interaction Studies*, vol. 6, no. 2, pp. 253–286, 2005.

[22] J. Schwartz, L. Boë, N. Valle, and C. Abry, "The Dispersion-Focalization theory of vowel systems," *Journal of Phonetics*, vol. 25, no. 3, pp. 255–286, 1997.

[23] P.-Y. Oudeyer, "The self-organization of speech sounds," *Journal of Theoretical Biology*, vol. 233, no. 3, pp. 435–449, Apr. 2005.

[24] ——, *Self-Organization in the Evolution of Speech*. Oxford University Press, USA, 2006.

[25] M. Schroeder, B. Atal, and J. Hall, *Frontiers of Speech Communication Research*. London Academic Press, 1979, ch. Objective measure of certain speech signal degradations based on masking properties of human auditory perception, pp. 217–229.

[26] C. Moulin-Frier, R. Laurent, P. Bessière, J.-L. Schwartz, and J. Diard, "Adverse conditions improve distinguishability of auditory, motor and perceptuo-motor theories of speech perception: an exploratory bayesian modeling study," *Language and Cognitive Processes*, vol. 27, no. 7–8, pp. 1240–1263, 2012.

[27] A. Baranes and P.-Y. Oudeyer, "The interaction of maturational constraints and intrinsic motivations in active motor development," in *IEEE International Conference on Development and Learning*, 2011.

[28] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]," http://www.praat.org/, 2012.

[29] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, Aug. 1980. [Online]. Available: http://dx.doi.org/10.1109/TASSP.1980.1163420